K- Means Clustering Techniques - A Review

Arshleen
Department of CSE, Chandigarh University, Gharaun

Abstract - The Clustering is a fundamental technique of categorizing data into various groups such that the data belonging to one group is similar to each other but they are very distinct from data existing in the other groups. K Means Clustering method involves division of any given data set into k number of clusters. This algorithm is also known as nearest neighbor clustering. K means clustering is the most imperative and basic technique for the investigation of various data sets. Beforehand, different endeavors have been done to enhance the execution of K-means algorithm. The result of enhanced k-means has given great performance in terms of small to medium sized data in comparison to the huge and substantial data. This paper reviews the various strategies and techniques used in the literature along with its benefits and shortcomings. The paper would also investigate the future possibilities of advancements in k means algorithm.

Keywords **-** Clustering, *K-Means, Nearest Neighbor, Data points.*

INTRODUCTION

With the passage of years a sharp growth is seen in the Internet usage. This increase in the web utilization results in producing heaps of data which is expanding as the years passes. The investigation of such data and grouping it into clusters is a difficult job. Further, issue lies in putting away and recovering the data. The scientists have evaluated that the data becomes two fold after every 20 months.

Nonetheless the crude information cannot be utilized straightforwardly. Its genuine worth is anticipated by separating data that is helpful for making decisions. The investigation of data was a manual procedure in many areas. But now the individuals are searching for figuring innovations to mechanize the procedure [1] as the measure of information control and investigation is going past human abilities.

1. CLUSTERING

The process of clustering plays an important role in the analysis and mining of data in various applications [2]. The data is divided into distinct classes on the basis of its attributes and qualities.
The clustering comes under the category of Unsupervised learning in which the expected outcome is not given during method of learning. Here, the data is clustered into groups on the basis of their statistical properties. Another category of clustering known as Supervised learning involves a trainer who gives input, and the desired output is generated. Many algorithms are formulated to accomplish clustering.

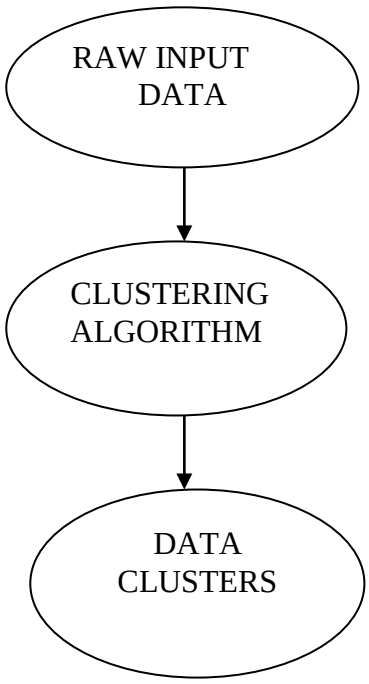

Figure1. Clustering Stages

There are two broad categories of clustering

namely hierarchical and partitioning clustering.

1. **Hierarchical Clustering**: It consists of clusters which are nested and arranged in the form of tree.
2. **Partitioning Clustering**: It consists of dividing various data into subsets so that each object of data is exactly in one subset.

This paper reviews the literature for the various procedures and techniques used for K Means clustering. It also discusses the benefits and drawbacks of the method along with the chances for the enhancement of K means algorithm.

3. K- MEANS CLUSTERING

Clustering is a vital and an imperative topic in the field of data mining which is effectively used in different applications. The process of clustering involves the division of data into distinct classes where each class of data has significant properties [3]. Therefore it can be concluded that classes comprise the objects having exactly similar attributes.

The K- Means clustering is the widely used clustering tool in the various fields of scientific and mechanical applications [4]. It is a strategy used for cluster investigation in which various observations are classified into k clusters and each of the observation belongs to the cluster with the nearest mean [5].

The K-Means algorithm is simple and easy to understand. The various steps are:

- Select the value of k as initial centroids.
- Repeat the following two steps for all the points in a set of data.
- Form k cluster by allocating every point to its most nearest centroid.
- Recalculate the centroid for each cluster until the centroid does not change.

The algorithm is used to a great extent in the field of data mining to extract useful data from the large sets of data.

4. LITERATURE SURVEY

K. A. Abdul Nazeer et al. [6] proposes k-implies calculation, for various arrangements of estimations of beginning centroids, produces distinctive groups. Last group quality in calculation relies upon the choice of starting centroids. Two stages incorporate into unique k implies calculation: first for deciding beginning centroids and second to allot information focuses to the closest bunches and afterward recalculating the grouping mean.

Soumi Ghosh et al. [7] present a relative talk of two bunching calculations in particular centroid based K-Means and agent question based FCM (Fluzzy C-Means) grouping calculations. This exchange is based on execution assessment of the effectiveness of bunching yield by applying these calculations.

Shafeeq et al. [8] present an altered K-implies calculation to enhance the group quality and to settle the ideal number of bunch. As info number of bunches (K) given to the K-implies calculation by the client. In any case, in the handy situation, it is exceptionally hard to settle the number of groups ahead of time. The technique proposed in this paper works for both the cases i.e. for known number of groups ahead of time and also obscure number of groups. The client has the adaptability either to settle the quantity of groups or information the base number of bunches required. The new bunch focuses are figured by the calculation by augmenting the group counter by one in every cycle until the point when it fulfills the legitimacy of bunch quality. This calculation will survive this issue by finding the ideal number of bunches on the run.

Junatao Wang et al. [9] propose an enhanced k means calculation utilizing commotion information channel in this paper. The inadequacies of the customary k-implies bunching calculation are overwhelmed by this proposed calculation. The calculation creates thickness based identification techniques in light of qualities of clamor information where the revelation and preparing ventures of the clamor information are added to the first calculation. By pre-handling the information to prohibit these clamor information before bunching information sets the group attachment of the bunching results is enhanced altogether and the effect of commotion information on k means calculation is diminished adequately and the grouping results are more exact.

Shi Na et al. [10] present the investigation of weaknesses of the standard k-implies calculation. As k means calculation needs to figure the separation between every datum protest and all bunch focuses in every emphasis.. An enhanced k-implies calculation proposed in this paper.

Table 1 Comparison among various existing Approaches and its Limitations

S.No.	Author	Methodology Used	Country of Research	Objectives	Limitations
1.	K. A. Abdul Nazeer et al	K-Means Algorithm	India	Presented an improved clustering method that calculates the initial centroid values and also effectively allocated the data points to the clusters. Enhanced the correctness of K-Means Algorithm.	The limitation of this algorithm lies on the fact that despite the distribution of the various data points, it is still required to give count of clusters as an input
2.	Soumi Ghosh et al.	K-Means Algorithm, Fuzzy C- Means Algorithm	India	Performs relative investigation of Fuzzy C Means and K means algorithm based on the criteria of time complexity. K- Means algorithm seems far better than Fuzzy C-Means.	The calculation time taken is more because of the fuzzy measurements.
3.	Shafeeq et al.	Modified K-Means Algorithm	India	The exact number of devised on the basis of run method of clustering. It works good for both familiar and non-familiar number of clusters.	The technique devised takes more time for calculation than k means in case of big data sets.
4.	Junatao Wang et al.	K-Means Algorithm	China	The updated algorithm produces less nose data as compared to the earlier researches	The noise impact is more in cluster forming.
5.	Shi Na et al	K-Means Algorithm	China	Enhances the speed and decreases the calculative complexity.	The algorithm used for the selection of the centroid is not very effective.

5. CONCLUSIONS

In this paper k-implies grouping strategies and strategy are checked on. K-implies being generally acclaimed among information researcher requires promote change in different area of calculation. The exceptions, void groups what's more, choosing centroid for datasets are as yet a testing errand. Thus different further research expected to center around these said issues. Table I. presents different methods and its restriction is available in proposed k means calculation. They require facilitate improvement due to increment of size of information starting at now. This paper has make an endeavor to survey a huge number of papers to manage the present calculation of k-implies. Present examination show that k means calculation can be upgraded by choosing centroid point fittingly.

REFERENCES

[1] E. A. Khadem, E. F. Nezhad, M. Sharifi, “Data Mining: Methods & Utilities”, Researcher 2013; 5(12):47-59. (ISSN: 1553-9865).

[2] Namrata S Gupta, Bijendra S Agrawal, Rajkumar M. Chauhan, Survey On Clustering Technique of Data Mining, American International Journal of Research in Science, Technology, Engineering & Mathematics, ISSN:2328-3491

[3] Malwinder Singh, Meenaksh bansal , A Survey on Various K Means algorithms for Clustering, IJCSNS International Journal of Computer Science and Network Security, VOL.15 No.6, June 2015.

[4] A. Saurabh, A. Naik, "Wireless sensor network based adaptive landmine detection algorithm, " 2011 3rd International Conference on Electronics Computer Technology (ICECT), vol.1, no., pp.220, 224, 8-10 April 2011

[5] Amandeep Kaur Mann, Navneet Kaur Mann, Review Paper On Clustering Techniques ,Global Journal Of Computer Science And Technology Software & Data Engineering, VOL. 13 ,201

[6] K. A. Abdul Nazeer, M. P. Sebastian, Improving the Accuracy and Efficiency of the k-means Clustering Algorithm, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K.

[7] Soumi Ghosh, Sanjay Kumar Dubey, Comparative Analysis of K-Means and Fuzzy C-Means Algorithms, International Journal of Advanced Computer Science and Applications, Vol. 4, No.4, 2013

[8] Shafeeq, A., Hareesha ,K., Dynamic Clustering of Data with Modified K-Means Algorithm, International Conference on Information and Computer Networks, vol. 27 ,2012

[9] Junatao Wang, XiaolongSu, An Improved K-means Clustering Algorithm, Communication Software and Networks (ICCSN), 2011 IEEE 3rd International Conference on 27 may,2011 (pp. 44-46)

[10] Shi Na, Liu Xumin, Guan Yong, Research on K-means Clustering Algorithm: An Improved K means Clustering Algorithm, Intelligent Information Technology and Security Informatics, 2010 IEEE Third International Symposium on 2-4 April, 2010(pp. 63-67)